

DIABETES PREDICTION USING MACHINE LEARNING
TECHNIQUES

Project Submitted for the partial fulfillment of the requirement for the
award of the degree of

Master of Computer Application

Dr. A.P.J Abdul Kalam Technical University, Lucknow

Supervised By:

Mrs. Shilpi Garg
(Assistant Professor)

Submitted By:

Rishika Gupta
Roll No: 1819414916



ACADEMIC SESSION

(2018-2020)

H.R INSTITUTE OF TECHNOLOGY

GHAZIABAD

TABLE OF CONTENTS

Chapter No.	Topics	Page No.
	Student Declaration	I
	Certificate from the Supervisor	II
	Acknowledgement	III
	Summary	IV
	List of Symbols and Acronyms	V
Chapter-1	Introduction	9 to 11
	1.1 General Introduction	
	1.2 Problem Statement	
	1.3 Design and research	
	1.4 Brief Description of the Solution Approach	
	1.5 Comparison of existing approaches to the problem framed	
Chapter-2	Literature Survey	12 to 14
	2.1 Summary of papers studied	
	2.2 Statical major report from different health organization	
	2.3 Effects of diabetes	

Chapter 3: Requirement Analysis and Solution Approach 15 to 18

3.1 Overall description of the project

3.2 sample and sampling data

3.2 Requirement Analysis

3.5 Solution Approach

Chapter-4 Modelling and Implementation Details 19 to 27

4.1 Design Diagrams

4.1.1 Class diagrams / Control Flow Diagrams

4.1.2 Activity diagrams

4.2 Implementation details

4.3 Finding of the study

4.3.1 Diabetes risk based on the age group.

4.3.2 Diabetes risk as a factor of plasma

4.3.3 Diabetes risk as a factor of serum insulin level

4.4 Screenshots

4.5 Risk Analysis

Chapter-5 Testing (Focus on Quality of Robustness) 28 to 31

5.1 Testing Plan

5.2 Component decomposition and type of testing required

5.3 Limitations of the solution

Chapter-6	Findings, Conclusion, and Future Work	32 to 35
------------------	--	----------

6.1 Findings

6.2 Performance measure

6.2 Conclusion

6.3 Future Work

Chapter-7	References	36 to 37
------------------	-------------------	----------

(I)

DECLARATION

We hereby declare that this submission is my/our own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature of the Student

Name of the Student Rishika Gupta

Roll Number 1819414916

Place Ghaziabad, Uttar Pradesh

Date

(II)

CERTIFICATE

This is to certify that the work titled “**DIABETES PREDICTION USING MACHINE LEARNING**” submitted by **Rishika Gupta** in partial fulfilment for the award of degree of **MCA** of **H.R Institute of Technology, Ghaziabad** has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor

Name of Supervisor MRS. SHILPI GARG

Designation Assistant Professor

Date

(III)

ACKNOWLEDGEMENT

We want to express our special thanks of gratitude to our teacher **MRS. SHIPI GARG**, who gave us the golden opportunity to do this wonderful project of **DIABETES PREDICTION USING MACHINE LEARNING**. She whole heartedly supported us at all the stages of the project and offered her invaluable experience and knowledge to guide us throughout the project and accomplish the project goals.

Further, we would like to thank our parents and friends who helped in finalizing this project within the limited time frame.

Signature of the Student

Name of the Student Rishika Gupta

Roll Number 1819414916

Date

(IV)

SUMMARY

Our project first analyses the pima diabetes dataset of india accuracies of existing Machine Learning techniques, including Logistic Regression, XGBoost, Random Forests, Support Vector Machine (SVM), Decision Tree, Gradient Descent Boosting, LightGBM and CatBoost. We have compared the accuracies of all these models and then proposed an algorithm that uses all these techniques to selectively remove features from the dataset so that the dataset is pruned and is left with only important features. The accuracies of all the models is again computed with the reduced feature set and is compared to the original feature set. Weka toolkit is used to compute the PCA and the Information Gain for every method. It is found that the accuracy improves with the reduced features for every technique as compared to the original feature set. Hence, it can be concluded that using our algorithm, patient who are about to diabetic can be more accurately predicted. XGBoost is then applied on these features and the resultant diabetes prediction accuracy is found to be higher than before. This is used to predict if a patient is diabetic or not.

Student Name :

Rishika Gupta

Student Signature:

.....

Name of Supervisor :

Mrs. Shilpi Garg

Signature of Supervisor:

.....

(V)

LIST OF ACRONYMS

1. RF: Random Forests.
2. SVM: Support Vector Machine.
3. LightGBM: Light Gradient Boosting.
4. CB: Cat Boost.
5. GB: Gradient Descent Boosting.
6. XGBoost: Extreme Gradient Boosting.
7. NLP: Natural Language Processing
8. RNN: Recurrent Neural Networks
9. LSTM: Long Short Term Memory

INTRODUCTION

1.1 General Introduction

Diabetes is a very common metabolic disease in all over world. Usually, type 2 diabetes happen in middle age and sometimes in old age. Diabetes is a disease caused due to the increase level of blood glucose. But nowadays, this disease are reported in children as well. There are multiple factors for developing diabetes like body weight, meal habits and sedentary life style. Undiagnosed diabetes can result in high blood gulcose level in blood referred to hyperglycemia which can cause complication such as diabetic retinopathy, nephropathy, neuropathy, cardiac stroke and foot ulcer. So, In early days detection of diabetes is very important for the health of the patient and enhancement of their immunity.

Now a days, diabetes is increasingly common in person's daily life. . This study compares the accuracy of advanced machine learning methods like Log-istic Regression, XGBoost, Random Forests, SVM, Decision Tree, Gradient Descent Boosting, LightGBM and CatBoost in predicting diabetes in patients and then make a design by comparing the accuracies of all the above techniques that aims to find the major causes of diabetes in all over the world.

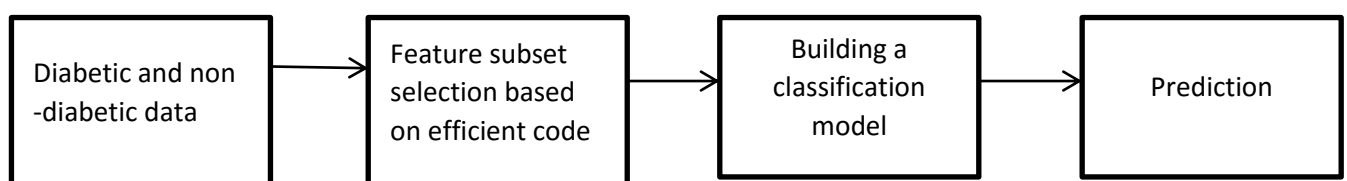
Random forest classifier is used to predict the diabetes using the pima diabetes dataset. It contains multiple factors on which we predict the diabetes like number of pragnences, Body mass index, Insulin, Age, Skin thickness .

1.2 Problem Statement

Diabetes is a common disease all over the world that affects a large majority of the people. Usually diabetes affects people after the age of 20. According to WHO statistics, the global prevalences of diabetes problem in adults above 18 years of age has increase to 8.5% in 2014. Diabetes prevalence has been rising more in mediate and low income countries. It also causes other illness like blindness, kidney failure, cholesterol and heart diseases. Due to diabetes and high blood glucose, death rate is also increases. In the early stage of Prediction of diabetes would help the patients to maintain the gulcose level under control. Data mining technique is very good in predictive analysis so this techniques is used to predict the risk of diabetes in the proposed approach. The performance of the algorithm is also improved and measured by feature selection and selection of training set.

1.3 Design of reseach:

The diabetes dataset is selected and divided into two – training and test dataset. A biggest problem is feature selection for knowledge discovery. The main aim is to select the feature subset that help in to produces higher classification accuracy. The classification algorithm is applied to make the classification model after selection of features. After that the model is applied to the test for predicting the diabetes risk. The performance metrica are evaluated and measure.



1.4 Brief Description of Solution Approach

In this study, we have firstly compared the accuracies of known advanced machine learning techniques like Log-istic Regression, XGBoost, Random Forests, SVM, Decision Tree, Gradient Descent Boosting, LightGBM and CatBoost in predicting diabetes and finest accuracy is taken into account.

An algorithm is proposed that selectively removes features from the dataset to find the most critical features responsible for diabetes in middle and lower age using various Machine Learning techniques.

The accuracy is then compared with the reduced features for every technique as compared to the original feature set to find whether the reduced feature set improved accuracy.

XGBoost is then applied on these features and the resultant diabetes prediction accuracy is found to be higher than before. This is used to predict if a customer is diabetic or not.

1.5 Comparison of existing approaches to the problem framed

After looking at all existing approaches, we found that in every approach, only a limited methods like one a common technique along with some boosting method or random forests or SVM are used. They have tested these techniques for the whole of the dataset and have shown which method gives the best accuracy. But feature selection and dataset reduction hasn't been applied.

Instead, we used the common algorithms along with the new algorithms, overall 8 methods have been used to get the best accuracy, and to get the best possible accuracy we have also reduced our dataset i.e. removed the columns that weren't useful. Our technique helped us to get the best possible accuracy.

LITERATURE SURVEY

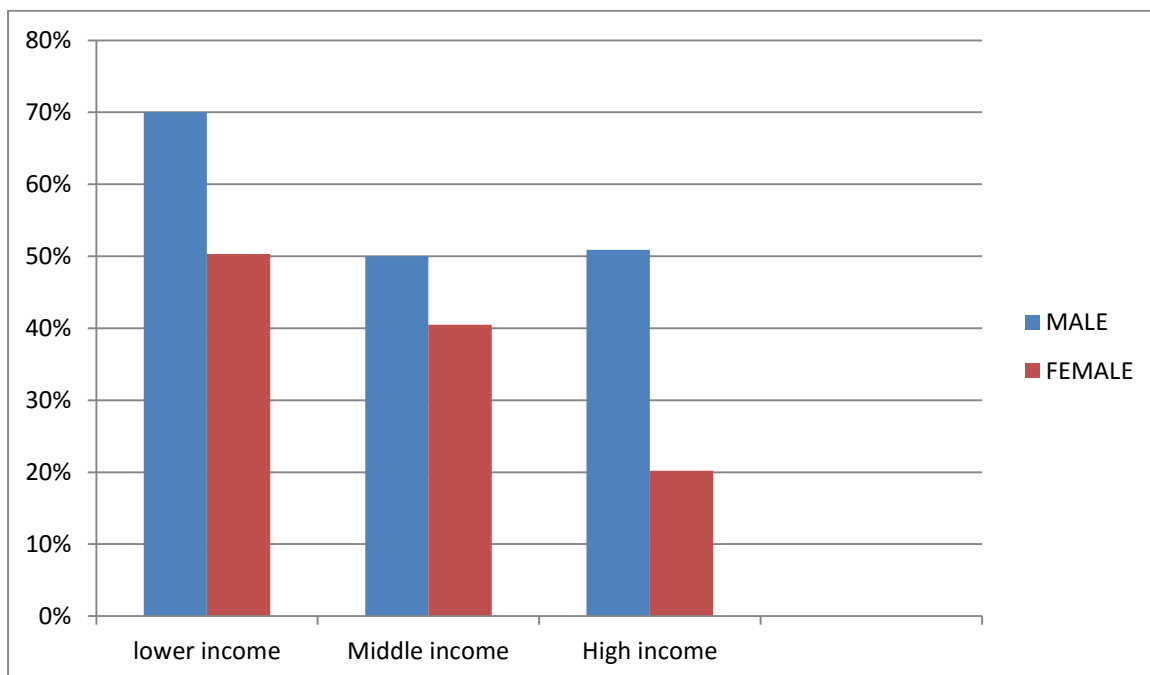
2.1 Summary of papers studied

General data mining methods are used to fit and predict the customer attrition models. While theory-driven statistical analyses would minimize the observation and variables, making it easier to find conclusions, however, data mining techniques has enhanced the capacity to lift large amount of data and several superior models can be tested at once.

Au (2003) found that the tree-based model is the most accurate amongst the three methods for predicting customer attrition as it uses the least number of predicting variables. Also, the logistical regression technique and the neural network technique, as opposed to the tree-based method which even includes missing data. However, data mining techniques has enhanced the talent to handle big amount of data and a lot of models can be tested at once. Three different methods of modelling are tested in this paper, namely, the logistic reg-based method, the random tree-based method, and the ANN method. Evaluation of these models are done on the basics of the receiver-operating characteristic (ROC) curve. The tree-based technique is found to be the most accurate among the three methods for predicting customer attrition because it uses the least number of variables predicting. The logistical regression method and the neural network method also use cases that do not have missing data, as opposed to the tree-based method that even includes missing data.

The yearly report of World Health Association, sum up the no. of individuals experiencing the disease of diabetes is 420 million the year. Consistently, there is a huge rising in the number individuals experiencing the disease of diabetes in different healing center. The world health organization (WHO) reports on “DIABETES CARES 2018” with the help of American Diabetes Association.

There is a graph for diverse individuals matured between 29 and 70 yrs, there levels passes because of hypertension.



Diabetes mellitus is chronic, it causes because of the high sugar level in the blood. It causes because of the inappropriate working of our pancreatic beta cells. It can affect on different part of the body which in corporates parcreas glitch, risk of heart ailment, kidney disappointment, nerve harm, meal issues, ketoacidosis, visual unsettling influences, and other eye issues and glaucoma etc.

There are various purposes behind this like a way of life of a man spend, the absence of activities, sustenance propensities, smoking, high cholesterol, high

blood pressure (Hyperglycaemia) etc. which fundamentally increment the risk of treating diabetic problem. It can effect any range of ages, including youngsters to grown-up and matured person.

Pancreas is an organ of our body that situated in midriff area. Pancreas has beta cell that discharge the insulin to the circulation system, to absorb the excessive sugar substance from the blood into liver. If sugar level become low then alpha cells help to maintain the sugar level in the blood.

2.2 STATICAL MAJOR REPORT FROM DIFFERENT HEALTH ORGANIZATION:

1. In 2017, the statistic diabetes report for Center Disease Control and Prevention (CDC), gives the facts that, United state has 30.3 million people has diabetes in which 23.1 are analysed and 72 billons are unidentified.
2. In 2018, the American Diabetes Association models gives a report about “Order and Finding diabetes “ which are not corporate to the arrangement of diabetes.
3. In 2017, Global gives details for diabetes by world wellbeing association, it shows the weight of diabetes and inconvenience of diabetes.

2.3 EFFECTS OF DIABETES:

- Loss of vision
- Kidney neuropathy
- Liver problem
- Heart problem
- Foot issues

REQUIREMENT ANALYSIS AND SOLUTION APPROACH

3.1 Overall description of the project

Our project's idea is to find the different aspects that causes the diabetes. There are many peoples that do not know that they are suffering from diabetes. When diabetes is not treated it could become dangerous for life. When diabetes detected earlier then it can reduce major complication like heart disease,, brain stroke etc. So with the help of this project, Any person can check that it has diabetes or not.

The first step is to define purpose of diabetes prediction. Once we collect enough data to make a prediction that is satisfactory, we can then start creating a predictive model. This also includes validation and tuning of the model to get from the dataset the most accurate diabetes prediction.

We then implemented eight classifiers, namely:

- Logistic Regression
- XGBoost
- Random Forests
- Support Vector Machine (SVM)
- Decision Tree
- Gradient Descent Boosting
- LightGBM
- CatBoost

to predict the behaviour of the patient by comparing the accuracy of these techniques. We have proposed an algorithm based on these techniques that takes

into account the best accuracy obtained and selectively removes features from the dataset to find the most critical features for diabetes prediction.

3.2 Sample and Sampling design :

The Pima Indian Diabetes dataset is chosen as the sample for the experimental setup. This dataset contain the record of diabetic and non-diabetic patients. It contain eight attributes and class attributes. There are 768 total instances available in the dataset. All the patients in the dataset are above 12 years of age and they are pima Indians. The attributes of dataset are shown in table below:

Attribute ID	Attribute Name	Attribute description
A1	Pregnant Time	Number of times pregnant
A2	Plasma glucose	Plasma glucose concentration a 2hr in an oral glucose tolerance test.
A3	Blood pressure	Diastolic Blood Pressure(mm hg)
A4	Skin thickness	Triceps Skin fold thickness(mm)
A5	Insulin	2-hour serum insulin(u/ml)
A6	BMI	Body Mass Index
A7	Pedigree	Diabetes Predigree Function
A8	Age	Age in years
A9	Class variables	Zero or one

The dataset is classified using random forest classifier algorithm and a model has been built. We were chosen 70% of the records to be the training set and the remaining 30% are taken as the test set.

3.3 Requirement Analysis

Functional Requirements:

- The system should be equipped with libraries such as Scikit, Pandas, XGBoost and Sklearn
- The algorithm should be able to provide better accuracy than existing algorithms to aid in better diabetes prediction.
- The algorithm should also considerably reduce the feature set and should only keep important features.

Non-functional Requirements:

- The algorithm should be optimal, and should not be a memory hog.
- The algorithm should be scalable and be able to run on different systems.
- The algorithm should not take a long time to run.

3.4 Solution Approach :

Random forest classifier is used to predict the data and predict the data with high accuracy.

Algorithm to predict the PIMA Diabetes dataset using the Random forest classifier.

Step 1: PIMA Diabetes data set is collected.

Step 2: Normalised all the value except the class variable which contain either “0” or “1”.

Step 3: After normalized the data set, we import all relevant packages.

Step 4: Data split for training and testing (70% and 30%).

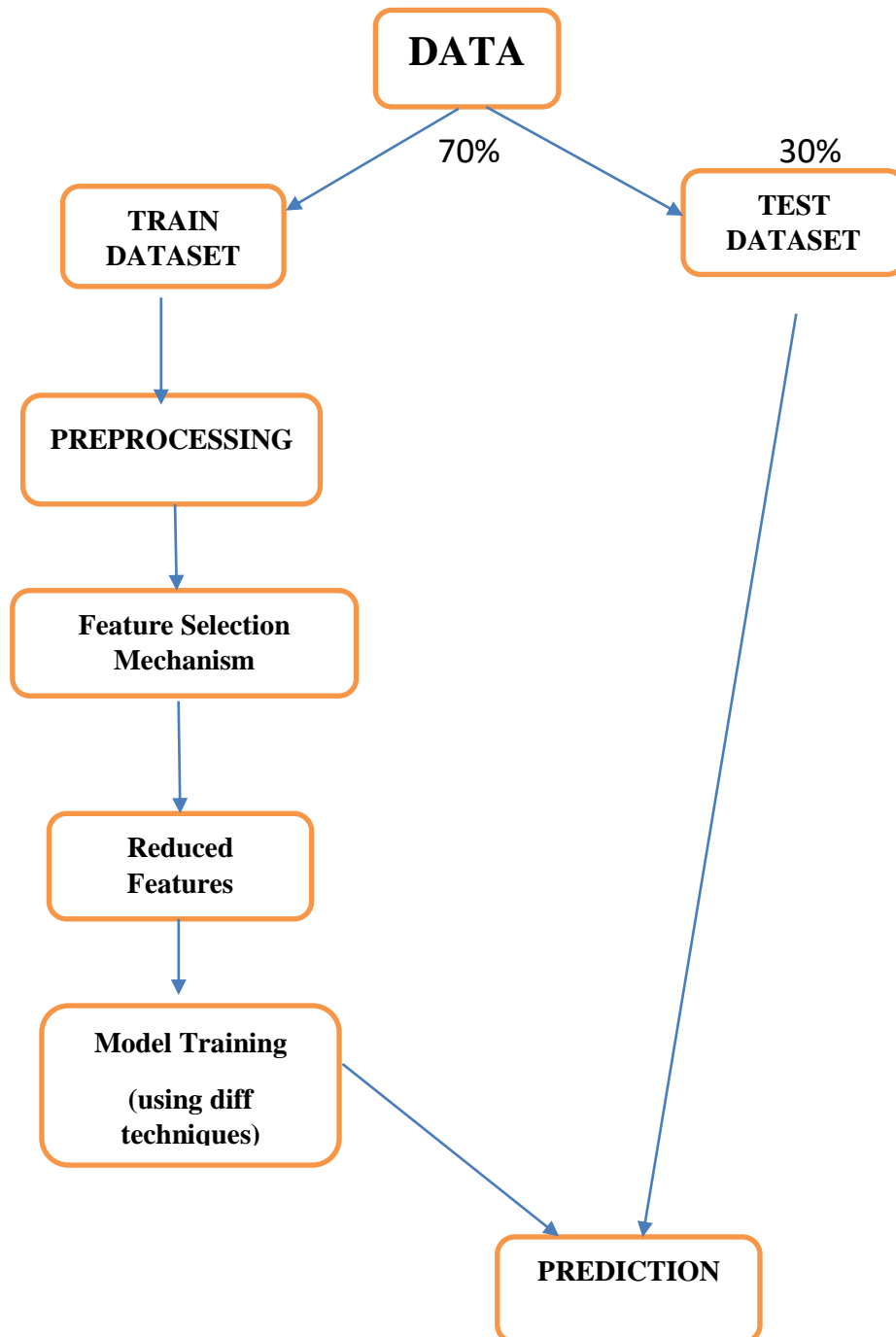
Step 5: Create the classifier model that we used in the dataset.

Step 6: Applying random forest classifier algorithm to train data and finding the accuracy of model testing data.

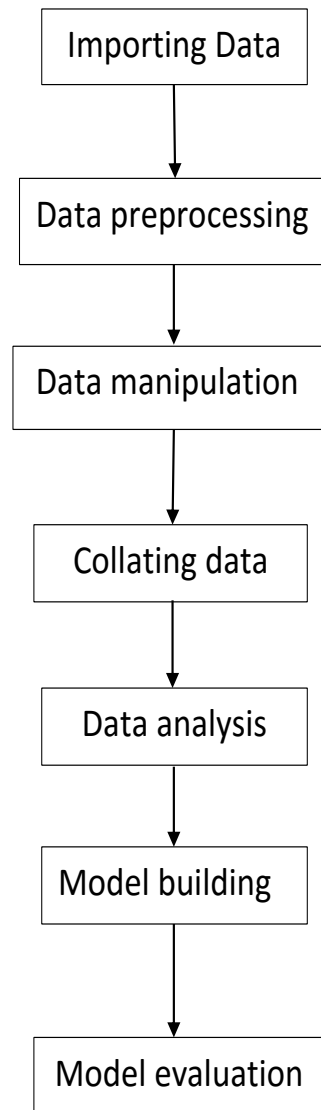
MODELLING AND IMPLEMENTATION DETAILS

4.1 Design Diagrams

4.1.1 Control Flow Diagram



4.1.2 Activity diagram



4.2 Implementation details

The project was done on Jupyter Notebook. For coding purpose, we used Python. All the algorithms we have used were coded in python and correspondingly accuracies of all the algorithms were found.

Various Machine Learning Algorithms that we used are:

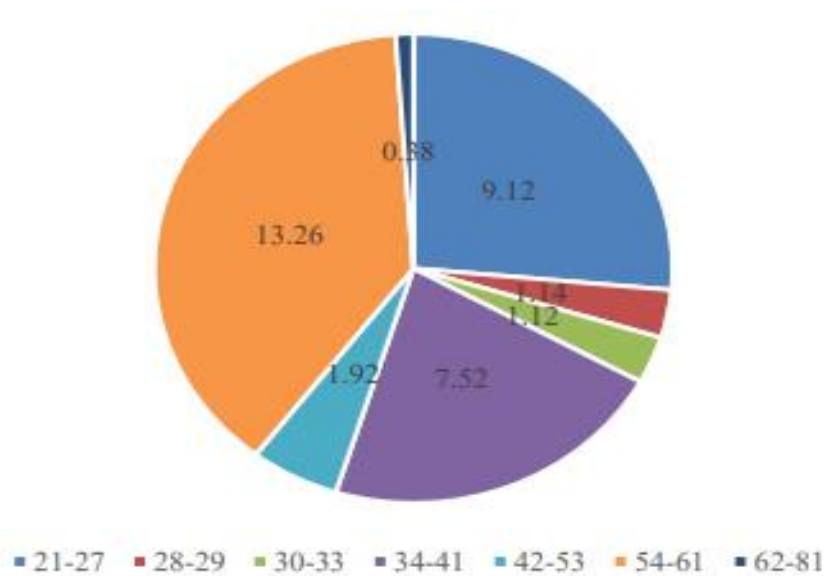
- *Logistic Regression* is used to define information and explain the connection between one dependent variable and one or more independent ordinal, nominal, interval features.
- *Random Forest*, which operates by constructing a large number of decision trees at the moment of training and outputting the class mode or mean prediction of the individual trees.
- *Support Vector Machine (SVM)* technique generates a template by assigning to one or the other category fresh instances, making it a non-probabilistic linear classification binary.
- *Extreme Gradient Boosting (XGBoost)* supports three primary gradient boosting types, Boosting Gradient, **Stochastic** Boosting Gradient, **Regularized** Boosting Gradient. **XGBoost increases execution speed and model performance.**
- *Decision Tree* is a tool that provides an outcome, that uses a tree-like decision technique and its potential implications, including fortuitous

outcomes. It is one way displaying an algorithm with only conditional statements of control.

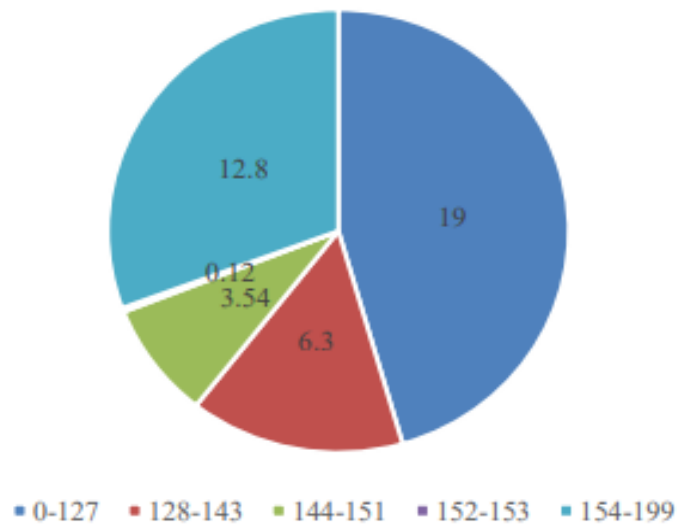
- ***LightGBM*** is a framework for boosting gradient using algorithms based on tree learning. It is intended for distribution and efficiency at higher training speed and faster efficiency, reduced storage usage and improved precision. It supports the learning of parallel and GPU and can handle large data.
- ***CatBoost*** operates with a multitude of information kinds to assist broad range of problems that companies presently face. It delivers excellent outcomes without compressive information preparation that other machine learning methods typically require.

4.3 Finding of the study:

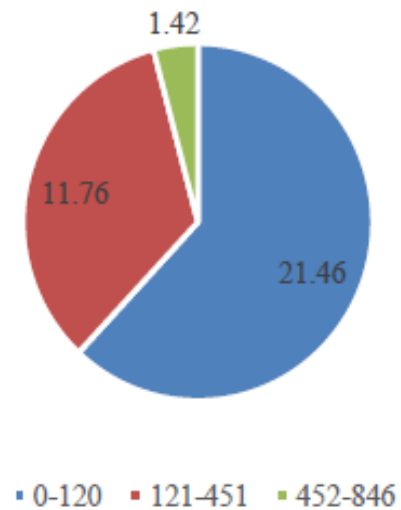
4.3.1 Diabetes Risk based on Age Groups :



4.3.2 Diabetes risk as a factor of plasma glucose level :



4.3.3 Diabetes Risk as a factor of Serum Insulin Levels :



Accuracies from all the eight algorithms have been noted and XGBoost has been observed to give our model the highest accuracy.

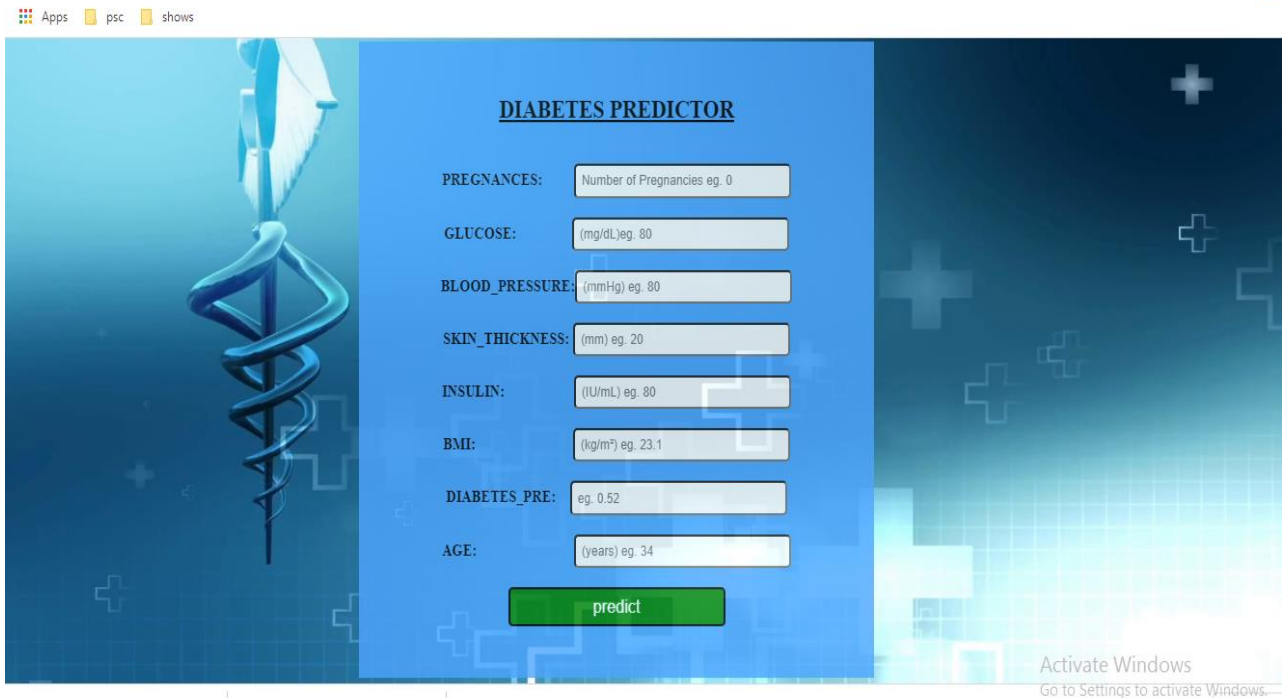
In our proposed algorithm, we have used feature scaling and feature selection techniques. Feature scaling is a technique for limiting the range of variables so they can be compared on common grounds. It is performed on continuous variables. We have used MinMax Scaler to limit the range of variable in 0-1.

We have also built a web app for data visualization purposes that represents the data on bar graphs and pie charts. This app was built using dc.js and d3.js, combined with Flask. It is an interactive app, where the data elements change when the user interacts with them. The data is dynamically changed if the user wants to view the data of a particular range of years or a specific gender of customers.

There is also a data prediction element in this web app. Using the features extracted in the above algorithm, the user is asked to input the variables of those specific features. Once all the features are input, and the user presses the 'Predict' button, the Flask code is called, which uses XGBoost on the selected features and presents the user with a score depicting whether the given user, based on the input features, is diabetes or not.

SCREENSHOTS

Apps psc shows



DIABETES PREDICTOR

PREGNANCES:

GLUCOSE:

BLOOD_PRESSURE:

SKIN_THICKNESS:

INSULIN:

BMI:

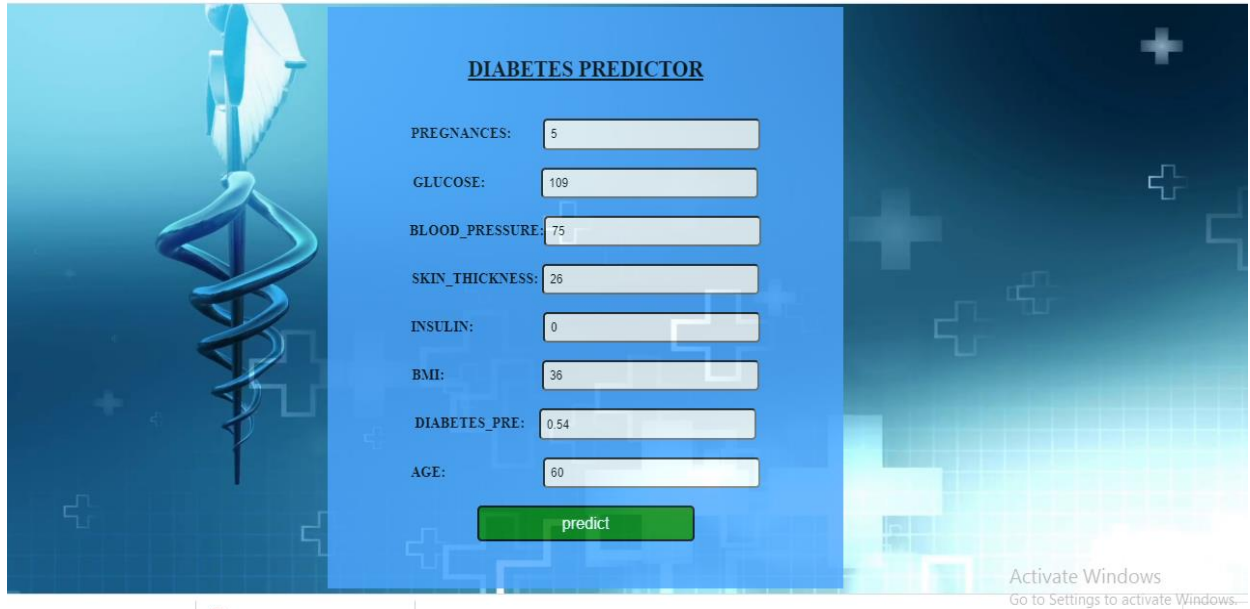
DIABETES_PRE:

AGE:

Activate Windows
Go to Settings to activate Windows.



Apps psc shows



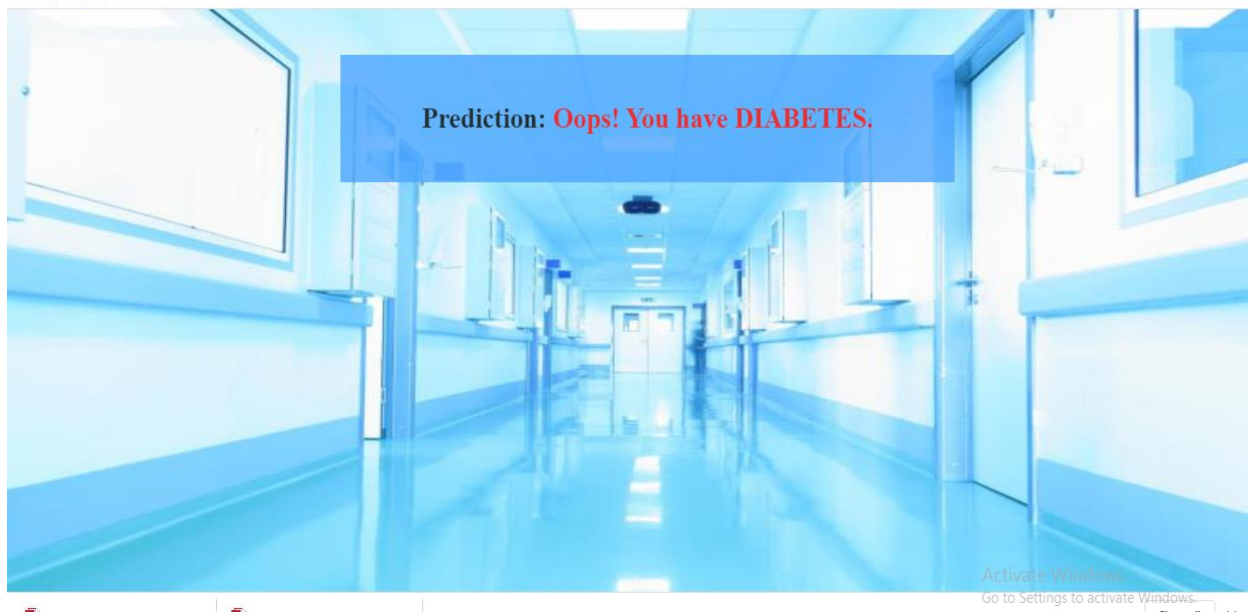
The image shows a web-based application titled "DIABETES PREDICTOR". On the left is a vertical blue bar with a white caduceus symbol. The main area has a light blue background with a grid of white plus signs. It contains several input fields for medical data, each with a label and a value: PREGNANCES (5), GLUCOSE (109), BLOOD_PRESSURE (75), SKIN_THICKNESS (26), INSULIN (0), BMI (36), DIABETES_PRE (0.54), and AGE (60). A green "predict" button is at the bottom. The right side of the interface is a dark blue vertical bar with white plus signs. At the bottom right, there is a watermark that says "Activate Windows Go to Settings to activate Windows..."

Parameter	Value
PREGNANCES	5
GLUCOSE	109
BLOOD_PRESSURE	75
SKIN_THICKNESS	26
INSULIN	0
BMI	36
DIABETES_PRE	0.54
AGE	60

predict

Activate Windows
Go to Settings to activate Windows...

Apps psc shows



4.4 Risk Analysis

Risk ID	Classification	Description of Risk	Risk Area	Probability	Impact	RE(P*I)
1	Development Environment – Development Process	Huge size of data	Data pre-processing or data cleaning	0.5	H	$0.2*5=1$
2	Development Environment – Development Process	Missing values of data	Data pre-processing or data cleaning	0.1	L	$0.1*1=1$
3	Development Environment – Development Process	Incorrect/missed database entry	Project Development	0.1	M	$0.1*3=3$
4	Program Constraints – Resources	Hardware limitation (resources)	Output fetching	0.4	M	

TESTING

5.1 Testing Plan

Type of Test	Will Test be performed?	Comments/Explanations	Software Development
Requirements testing	Yes	Performed to check the various requirements from aspects like input, development and testing	Python, Excel, JavaScript, HTML, Java
Unit testing	Yes	Performed on individual modules of implementation	Python, JavaScript, HTML, Java
Integration testing	Yes	Performed after merging different modules in Unit testing	Python, JavaScript, HTML, Java
Performance testing	Yes	Performed to check the responsiveness of our implementation under different work load	Python, JavaScript, HTML, Java
Load testing	Yes	To check the efficiency and optimal nature of our implementation	Python, JavaScript, HTML, Java

Activity	Start Date	Completion Date	Hours	Comments
Obtain pre-processed input data	15/01/2019	27/01/2019	Up to 2 hours	Raw data when put to cleaning took time due to heavy size
Training and testing dataset and obtaining of Word Cloud	28/01/2019	15/02/2019	10-15 hours	Was time consuming as the algorithms had a long run time
Feature selection and accuracy comparison	16/02/2019	05/03/2019	7-8 hours	Selectively removed features to tune the accuracy
Development of the web app and Android app	06/03/2019	12/05/2019	10-15 hours	Was time consuming since both apps had to be built from scratch

Test Environment

1. **Jupyter Notebook:** The Jupyter Notebook App is a server-client application that allows editing and running notebook document via a web browser.
2. **Weka:** Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.
3. **Dc.js:** dc.js is a javascript charting library with native [crossfilter](#) support, allowing highly efficient exploration on large multi-dimensional datasets
4. **Android Studio:** Android Studio is Google's official integrated development environment, built on JetBrains ' IntelliJ IDEA software and specifically designed for Android development.

Software Requirements

1. **Operating System:** Windows 10
2. **Languages:** Python, JavaScript, HTML, CSS
3. **Tools:** Jupyter Notebook, Weka

Hardware Requirements

1. **CPU:** Intel Core i5 2.4GHz
2. **Computer memory (RAM):** 8GB
3. **Hard disk:** 1TB

5.2 Component decomposition and type of testing required

S. No.	Components that require testing	Type of testing required
1	Obtain pre-processed input data	Requirements testing
2	Training and testing dataset on various algorithms	Unit testing
3	Smooth functioning of the website developed	Unit testing
4	Feature selection and accuracy comparison	Performance testing

5.3 Limitations of the solution

1. Type of data: Our solution only works on data that is textual and fails to work on data that is numerical.
2. Time taking: Our algorithm takes a long time to run and select features as it has to compute the accuracies of all 8 machine learning techniques.
3. Infrastructural constraint: Our algorithm requires more resources than currently available with us and hence takes a long time to process a large dataset of items.

FINDINGS, CONCLUSION AND FUTURE WORK

6.1 Findings

ALGORITHMS USED	ORIGINAL DATASET	PCA	INFO GAIN	REDUCED DATASET
LOGISTIC REGRESSION	0.8075	0.79	0.77	0.8175
DECISION TREE	0.74	0.75	0.74	0.7566
RANDOM FOREST	0.801	0.79	0.78	0.8109
SUPPORT VECTOR MACHINE	0.8143	0.80	0.79	0.8156
XGBOOST	0.81	-	-	0.8199
LIGHT GRADIENT BOOSTING	0.82	-	-	0.8118
CAT BOOST	0.74	-	-	0.8189
GRADIENT BOOSTING	0.81	-	-	0.8137

6.2 Performance Measure :

The performance of the model can be evaluated using different performance metrics. This report measures the performance of the algorithm using three performance metrics name as - accuracy, sensitivity and specificity. These measure are calculated by Confusion Matrix. Confusion matrix is a kind of table that is used to predict the performance of a classification model on a sample data. This matrix shows the number of True Positive(TP), False Negative(FN), False Positive(FP), True Negative(TF).

Format of confusion matrix:

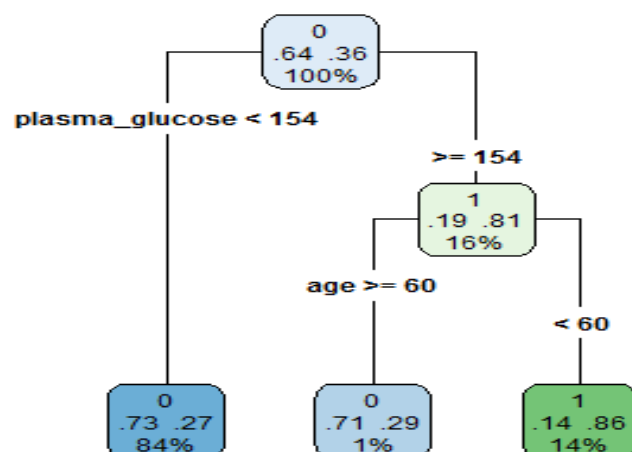
Actual vs Prediction	Positive	Negative
POSITIVE	TP	FN
NEGATIVE	FP	TN

The formulae for calculating the performance metrics are shown in equations

$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN)$$

$$\text{Sensitivity} = TP/(TP+FN)$$

$$\text{Specificity} = TN/(TN+FP)$$



6.3 Conclusion

In this research, the main objective is to find a model that predicts Diabetes in person when gives input and it provide higher accuracy rate then the existing models. In this model we compare different models, multiple classification algorithms and clustering algorithms were used and implemented. We can say that IDPA has highest accuracy when compared with other models and other researches models.

This model when included in real time apps(applications) in health sector can be used to predict diabetes with highest accuracy. The model has been trained to classify the diabetes patients from non-diabetes persons and it is used to predict the risk of diabetes on another dataset. The performance of the model has been evaluated using the performance measures such as accuracy, sensitivity and specificity.

The model can be enhanced by using real time dataset or hospital patient's data. It would be more beneficial if any user gets a mobile app which can predicts diabetes or non-diabetes and also manage to stores the patient information.

After applying all the above algorithms, and integrating with our own algorithm we noted the accuracies and calculated the PCA and Information Gain for WEKA for Logistic Regression, Decision Tree, Random Forest and Support Vector Machine Algorithms. Additionally, in our quest to find the best accuracy, we have managed to reduce the features of the dataset from 42 original features to 26 features. The accuracy improves with the reduced features for every technique as compared to the original feature set. Hence, it can be concluded that using our algorithm, customers who are about to churn can be more accurately predicted.

6.3 Future Work

The following areas will be explored to further improve upon the proposed algorithm-

1. To improve our accuracy even further, we can employ Deep Learning techniques like Recurrent Neural Networks (RNN), instead of just using Machine Learning techniques.
2. We can use Time Series Forecasting using Long Short Term Memory (LSTM) to improve the prediction.

REFERENCES

1. J.Tuomilehto, “Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance”, In proceedings of International Journal of Medical Research, vol. 344,no. 18,pp. 1343-1350, 2001
2. <https://www.who.int/news-room/fact-sheets/detail/diabetes>
3. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF). Retrieved 17 December 2008.
4. [http://archive.ics.uci.edu/ml/datasets/PimaIndiansDiabetes](http://archive.ics.uci.edu/ml/datasets/Pima%20Indians%20Diabetes)
5. K.C. Tan, E.J. Teoh, Q. Yua, K.C. Goh, “A hybrid evolutionary algorithm for attribute selection in data mining”, 2008 Published by Elsevier Ltd.
6. S.Vijayarani, S.Sudha, “Disease Prediction in Data Mining Technique – A Survey”, International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013.
7. Srideivanai Nagarajan,R.M ChandraSekaran,Data Mining Techniques for Performance Evaluation of Diagnosis in gestational Diabetes,IJCRAR 2014 pp91-98
8. Aiswarya Iyer,S.Jeyalatha Diagnosis of Diabetes Using Classification Mining Techniques,IJDKP 2015
9. Han Wu,Shengqi Yang,Type 2 Diabetes Prediction Model Based on Data Mining,Informatics in medicine unlocked 2018
10. Humar K, Novruz A. Design of a hybrid system for the diabetes and heart diseases.Expert Syst Appl 2008;35:82–9.
11. Patil BM. Hybrid prediction model for Type-2 diabetic patients. Expert Syst Appl 2010;37:8102–8108.

- 12.K.Rajesh,V.Sangeetha.Application of Data Mining Methods and Techniques for Diabetes Diagnosis.IJEIT 2012 , Volume 2 Issue 3.
- 13.Abdullah A.Alijumah,Mohd Gulam Ahmad,Application of data mining:Diabetes health care in young and old patients,Journal of King Saud University-Computer and Information Sciences 2013 25,127-136
14. Srideivanai Nagarajan,R.M ChandraSekaran,Data Mining Techniques for Performance Evaluation of Diagnosis in gestational Diabetes,IJCRAR 2014 pp91-98
- 15.Patil BM. Hybrid prediction model for Type-2 diabetic patients. Expert Syst Appl 2010;37:8102–8108.
- 16.Ahmad Aliza, MustaphaH Aida. Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus. ICDIPC 2011,Part I. CCIS 188; 2011. p. 537–45.
17. Marcano-Cedeño Alexis, Torres Joaquín, Andina Diego. A prediction model to diabetes using artificial metaplasticity. IWINAC 2011, Part II. LNCS 6687; 2011.p. 418–25.